

PB86-140639

DRAFT USER'S GUIDE FOR SFILES

General Software Corporation
Landover, Maryland

28 Aug 84



U.S. Department of Commerce
National Technical Information Service
NTIS

DRAFT
USER'S GUIDE FOR SFILES

Prepared for:

U.S. ENVIRONMENTAL PROTECTION AGENCY
OFFICE OF PESTICIDES AND TOXIC SUBSTANCES
EXPOSURE EVALUATION DIVISION
Task No. 1
Contract No. 68-02-3970
Russell Kinerson - Project Officer

Prepared by:

GENERAL SOFTWARE CORPORATION
8401 Corporate Drive
Landover, Maryland 20785

Jon Chen - Program Manager
James Pilotte - Task Leader

Submitted: August 28, 1984

BIBLIOGRAPHIC INFORMATION

PB86-140639

SFILES. User's Guide (Draft).

28 Aug 84

PERFORMER: General Software Corp., Landover, MD.
Contract EPA-68-02-3970

SPONSOR: Environmental Protection Agency, Washington, DC.
Office of Pesticides and Toxic Substances.

For system on magnetic tape, see PB86-140621. Sponsored by Environmental Protection Agency, Washington, DC. Office of Pesticides and Toxic Substances.

SFILES is a program for entering molecular structures and storing them in a connection table form on a disk file for later use in Quantitative Structure Activity Relationship studies (QSAR). The user may enter a structure by drawing its hydrogen suppressed representation on a TEKTRONIX 4014 or compatible graphics terminal; or by building the connection table via one of two terminal independent structure entry systems. The graphical method of structure entry is the default method used in the program, and is the method which allows the user the most versatility in structure entry. Use of the graphical method is described in the main part of this user's guide. The two terminal independent structure entry methods, SMILES and SENTRY, are described in Appendices A and B respectively of this user's guide. Implementation of the SENTRY method is not yet complete. It is anticipated that the implementation will be completed and Appendix B available by December 1984.

KEYWORDS: *Molecular structure, *Environmental surveys,
*Quantitative structure activity relationship,
*SFILES programming language, *User
manuals(Computer programs).

Available from the National Technical Information Service,
SPRINGFIELD, VA. 22161

PRICE CODE: PC A03/MF A01

TABLE OF CONTENTS

1. Introduction	1
2. Examples	1
2.1 Example 1 - How to add and review a simple structure	1
2.2 Example 2 - How to build a more complex structure from a simple structure, alter a structure, and recover from errors	8
3. General Information	17
3.1 Checking file contents	17
3.2 Adding and changing structures	21
3.2.1 Alterations	21
3.2.2 Entry of charge information	22
3.2.3 Entry of ring information	22
3.2.4 Specifications of stereochemistry	23
3.2.5 Drawings structures containing double bonds to insure correct seometrical relationships	24
3.3 Symbols for atom/bond types	26
3.4 Reasons for bell ringings	26
4.0 Summary of SFILES command	27

Appendix A

1. INTRODUCTION

SFILES is a program for entering molecular structures and storing them in a connection table form on a disk file for later use in Quantitative Structure Activity Relationship studies (QSAR). The user may enter a structure by drawing its hydrogen suppressed representation on a TEKTRONIX 4014 or compatible graphics terminal; or by building the connection table via one of two terminal independent structure entry systems. The graphical method of structure entry is the default method used in the program, and is the method which allows the user the most versatility in structure entry. Use of the graphical method is described in the main part of this user's guide. The two terminal independent structure entry methods, SMILES and SENTRY, are described in Appendices A and B respectively of this user's guide. Implementation of the SENTRY method is not yet complete. It is anticipated that the implementation will be completed and Appendix B available by December 1984.

User interaction with SFILES is accomplished via entry of an SFILES command (summarized in Section 4) in response to the program prompt "TASK?". The two examples below illustrate the SFILES graphical entry method as well as the general operation of the program.

2. EXAMPLES

The purpose of the following examples is to help you become familiar with basic SFILES commands. It is recommended that you sign on to the TEKTRONIX terminal and go through the examples step-by-step.

2.1 Example 1

Example 1 shows how to

- (1) access SFILES,
- (2) add a structure to the backbone disk files,
- (3) review the structure after it has been entered.

SFILES is an estimation procedure in the Graphical Exposure Modeling System, GEMS, and is accessed by entering E in response to the GEMS prompt: "Enter an operation:" and by entering SFILES in response to the prompt "Enter an estimation procedure:". Once in SFILES, the message "THIS IS SFILES" is presented on the terminal. The first prompt in SFILES is:

ENTER STRUCTURE TYPE:

the possible responses to this prompt are STRUCTURE, SUBSTRUCTURE, or BACKBONE. For this example the BACKBONE structured type is selected. This signals that subsequent commands refer to a compound that is classified as a backbone structural type. In example 2 we will retrieve this backbone structure to build a compound of type STRUCTURE.

You now are prompted to specify a task. In response to this prompt, you can type HELP to get the list of possible commands. Now type ADD in response to the TASK? prompt.

The next message you see is

START WITH BACKBONE (B), STRUCTURE (S), NOTHING (N) ?

For this example, type N to start with nothing. At this point the screen clears, the cross-hairs and the list of SFILES encoding directives appear. The program is ready to accept encoding information.

The structure you will enter with the instructions below is pictured in Figure 1. In the following list of encoding steps, the abbreviations <r> and sp stand for the RETURN key and the space bar, respectively.

<u>Cursor Location</u>	<u>Keys Entered</u>	<u>Screen Changes</u>
1. position of atom 1	N <r>	N appears
2. position of atom 2	sp <r>	C appears, cross-hairs disappear
3.	4 <r>	aromatic bond (dashed line) drawn between atom 1 and atom 2; cross-hairs reappear
4. leave on atom 2	sp <r>	
5. position of atom 3	sp <r>	C appears; cross-hairs disappear
6.	4 <r>	aromatic bond between 2 and 3; cross-hairs reappear
7. leave on atom 3	sp <r>	
8. position of atom 4	sp <r>	C appears; cross-hairs disappear
9.	4 <r>	aromatic bond between atoms 3 and 4
10. leave on atom 4	sp <r>	

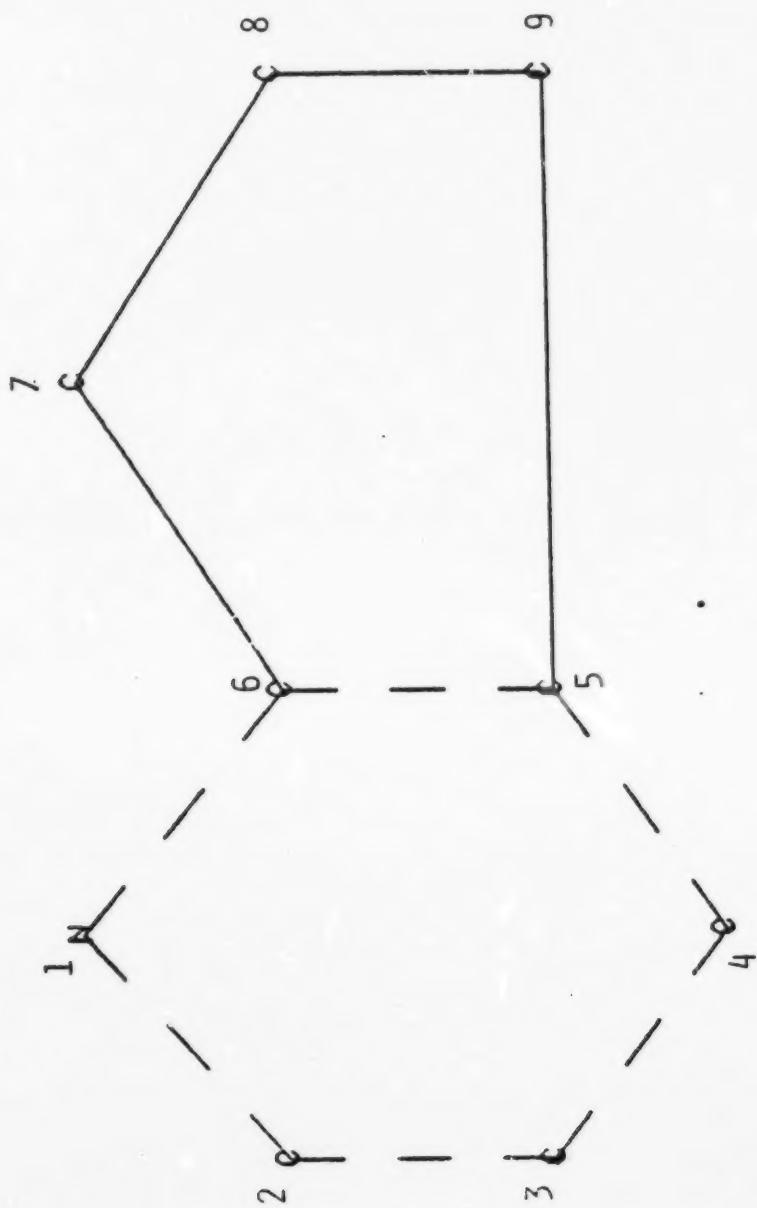


Figure 1. Structure Used in Example 1

11.	position of atom 5	sp <r>	C appears; cross-hairs disappear
12.		4 <r>	aromatic bond between atoms 4 and 5
13.	leave on atom 5	sp <r>	
14.	position of atom 6	sp <r>	C appears; cross-hairs disappear
15.		4 <r>	aromatic bond between atoms 5 and 6
16.	leave on atom 6	sp <r>	
17.	position of atom 1	sp <r>	cross-hairs disappear
18.		4 <r>	aromatic bond between atoms 6 and 1; first ring is drawn
19.	position of atom 6	sp <r>	
20.	position of atom 7	sp <r>	C appears; cross-hairs disappear
21.		1 <r>	single bond between atoms 6 and 7
22.	leave on atom 7	sp <r>	
23.	position of atom 8	sp <r>	C appears; cross-hairs disappear
24.		1 <r>	single bond between atoms 7 and 8
25.	leave on atom 8	sp <r>	
26.	position of atom 9	sp <r>	C appears; cross-hairs disappear
27.		1 <r>	single bond between atoms 8 and 9
28.	leave on atom 9	sp <r>	
29.	position of atom 5	sp <r>	cross-hairs disappear
30.		1 <r>	single bond between atoms 9 and 5

The screen should look like Figure 2 at this point. We will now add this structure to the backbone disk files. To exit the encoding section of SFILES, type D. The message

ENTER DIRECTIVE

appears. Type 5 to specify the FINISH directive. An asterisk marks the directive you've entered and numbers appear to the left of each atom in the structure. These numbers are produced by the ring section of SFILES which is automatically executed when a new structure is entered. Each number denotes the number of ring structures in which the corresponding atom is a member.

S-FINISH

4-ALTER

3-STEREO

2-RING

1-CONNECTION

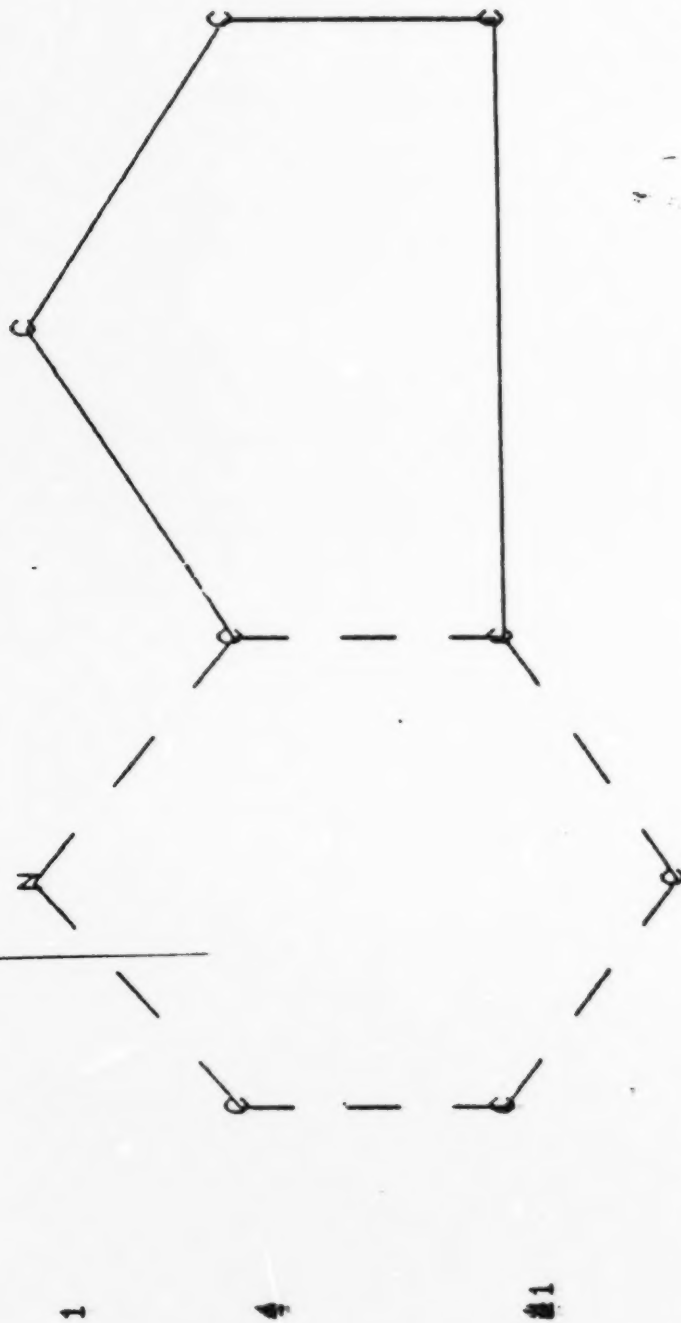


Figure 2. Structure Entered in Example 1

Now type <r> to get the next set of prompts. The title and label you enter will appear as identifying information on output you may want to generate later. The library access number (Direct Access Number, DAN) serves as an address in the disk file. Although the program informs you of the next sequentially available DAN, you can select a number randomly to enter in response to the prompt

INPUT LIBRARY ACCESS NUMBER

If this location already contains a structure, the program will ask you to choose another number - it won't overwrite the structure already there. (To wipe out a structure that you no longer want to keep in the file, you must use the DELETE command, which is described in example 2). If you choose not to save the structure you may enter a DAN of zero.

You are now asked whether or not you want to add another structure. Type NO or N to return to the full task list. You can review the structure you've just entered by typing REVIEW in response to the TASK? prompt, and then enter the appropriate library access number. If you respond to the prompt

WANT TO CHANGE DRAW OPTIONS?

with no, the default drawing and labelling options are assumed and the reviewed structure will look like Figure 3. Notice that the carbon atoms are not labelled.

After the structure has been drawn, type <r> to get the message

PRODUCE PRINTED VERSION?

If you type YES or Y in response, information about the structure will be written to a file called SFOUT.DAT. You can examine this file at the terminal after exiting SFILES by typing

TYPE SFOUT.DAT

or you can generate a line-printer copy of the file by typing

```

*****
EXAMPLE 1
*****
FORMULA:  C8 H9 N
C: 8  O: 0  N: 1  S: 0  P: 0
SINGLE:  4  DOUBLE: 0  TRIPLE: 0
DELOCALIZED: 0
*****
DAN: 19  LABEL:
*****
M.W. 119.17
TOTAL: 9
AROMATIC: 6
*****

```

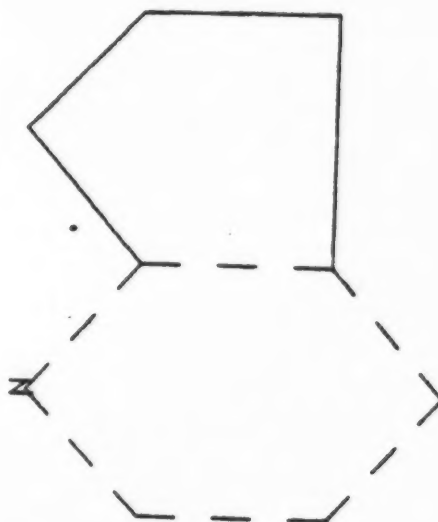


Figure 3. Review of structure in Example 3.

PRINT SFOUT.DAT

Figure 4 is a sample of the printed version output.

To test other draw options, copy the following sequence of responses to the prompts

```
REVIEW ANOTHER STRUCTURE ?  
Y  
INPUT LIBRARY ACCESS NUMBER ?  
<access number>  
WANT TO CHANGE DRAW OPTION ?  
Y  
SET DEFAULT VALUES ?  
N  
LABEL ATOM TYPE ?  
Y  
NUMBER ATOMS AS IN CONNECTION TABLE ?  
Y  
FULL SCREEN FOR DRAWING ?  
Y  
LABEL RE RING INFORMATION ?  
N  
SUPPRESS AUXILLARY INFORMATION ?  
Y  
BACK THRU OPTIONS AGAIN ?  
N
```

The drawing of these commands appear in Figure 5. After the structure has been drawn type <r>, then type N to return to the main command list. If you are ready to quit, type DONE to exit the program. If not, continue to example 2.

2.2 Example 2

Example 2 uses the structure entered in example 1 to build a more complex structure which is added to the structure files. Example 2 also shows how to enter stereochemistry information and how to correct errors made when entering a structure.

If you are continuing from example 1 (i.e. currently running SFILES) you should type ALTER in response to the TASK? prompt and then select structure type STRUCTURE. If you are just entering SFILES, select

```

*****
EXAMPLE 1
*****
FORMULA: C8 H9 N
C: 8 O: 0 N: 1 S: 0 P: 0
SINGLE: 4 DOUBLE: 0 TRIPLE: 0
DELOCALIZED: 0 TOTAL: 10
*****
M.W. 119.17
DAN: 19 LABEL:
TOTAL: 9
AROMATIC: 6
*****
ATOM SYMBOL | BOND FROM TO TYPE | ATOM IN RING(1:5)/SIZE
1 N | 1 | 1 2 3 4 5 6 7 8 9 | 6 6 6 6 6 6 5 5 5
2 C | 2 | 1 2 3 4 5 6 7 8 9 | 6 6 6 6 6 6 5 5 5
3 C | 3 | 1 2 3 4 5 6 7 8 9 | 6 6 6 6 6 6 5 5 5
4 C | 4 | 1 2 3 4 5 6 7 8 9 | 6 6 6 6 6 6 5 5 5
5 C | 5 | 1 2 3 4 5 6 7 8 9 | 6 6 6 6 6 6 5 5 5
6 C | 6 | 1 2 3 4 5 6 7 8 9 | 6 6 6 6 6 6 5 5 5
7 C | 7 | 1 2 3 4 5 6 7 8 9 | 6 6 6 6 6 6 5 5 5
8 C | 8 | 1 2 3 4 5 6 7 8 9 | 6 6 6 6 6 6 5 5 5
9 C | 9 | 1 2 3 4 5 6 7 8 9 | 6 6 6 6 6 6 5 5 5

```

Figure 4. Printed Connection Table.

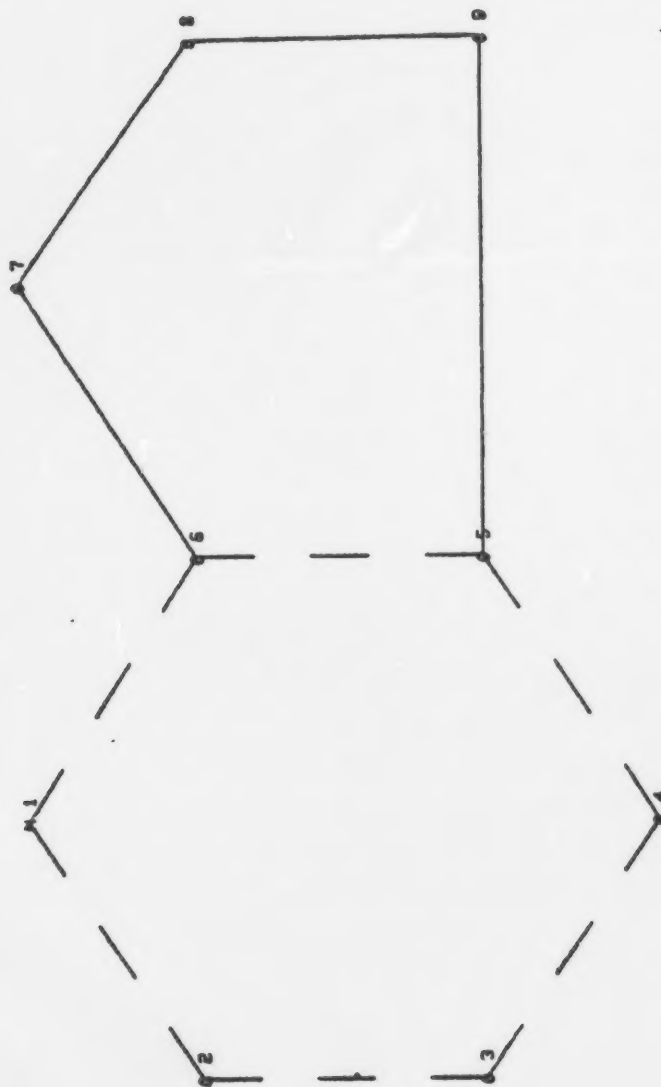


Figure 5. Alternate Review of Structure in Example 1.

STRUCTURE in response to the first prompt. In either case, enter the ADD directive in response to the next TASK? prompt.

Since you will use the backbone structure from example 1 to build a structure, type BACKBONE when you are asked what you want to start with, and then enter the access number you selected in example 1. Next, you should request to have the structure redrawn. At this point the backbone structure appears on the screen. Notice that the carbon atoms are not labelled and that the structure is drawn in large scale.

You are now in the ADD section of the SFILES program and must enter an encoding directive. Since you are going to attach new atoms and bonds to the current structure, you should type 1 for the connection directive. The cross-hairs appear and the program is ready to accept new encoding information. Refer to Figure 6 when following the encoding steps below.

Encoding steps:

	<u>Cursor Location</u>	<u>Keys Entered</u>	<u>Screen Changes</u>
1.	position of atom 9	sp <r>	
2.	position of atom 10	sp <r>	C appears; cross-hairs disappear
3.		1 <r>	single bond between atoms 9 and 10
4.	position of atom 10	sp <r>	
5.	position of atom 11	sp <r>	C appears; cross-hairs disappear
6.		1 <r>	single bond between atoms 9 and 11

At this point the screen may seem to be getting somewhat crowded at the right margin where atom 12 is going to be entered. To redraw the structure in smaller scale, you can type D for directive, then type @. To get back into ADD mode, type 1 for the connection directive and go to step 7.

7.	position of atom 11	sp <r>	
8.	position of atom 12	sp <r>	C appears; cross-hairs disappear
9.		2 <r>	double bond drawn between atoms 11 and 12
10.	position of atom 4	sp <r>	
11.	position of atom 13	F <r>	FL appears; cross-hairs disappear

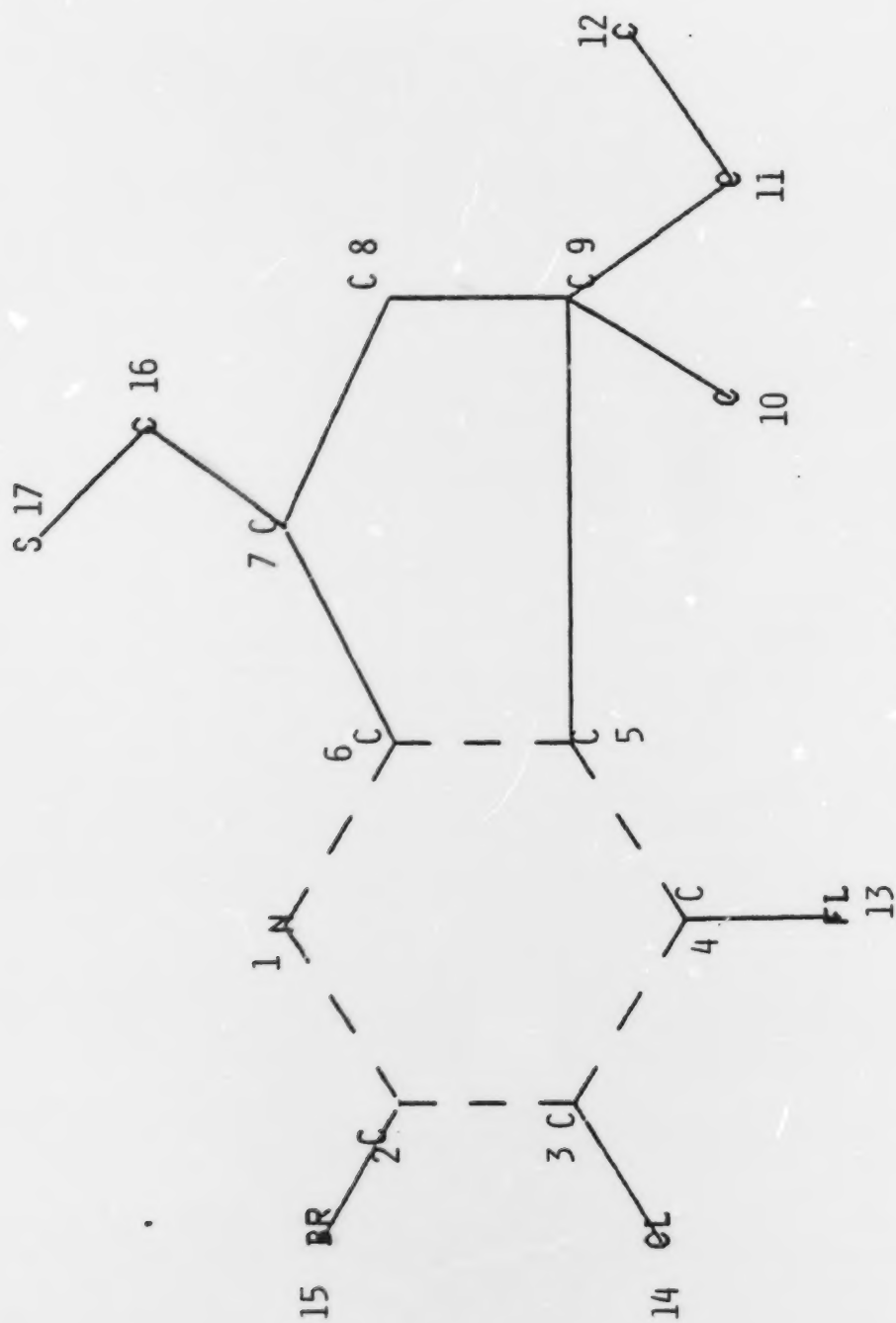


Figure 6. Structure Used in Example 2

12.	1 <r>	single bond between atoms 4 and 13
13. position of atom 3	sp <r>	
14. position of atom 14	C <r>	CL appears; cross-hairs disappear
15.	1 <r>	single bond between atoms 3 and 14
16. position of atom 2	sp <r>	
17. position of atom 15	B <r>	BR appears; cross-hairs disappear
18.	1 <r>	single bond between atoms 2 and 15
19. position of atom 7	sp <r>	
20. position of atom 16	sp <r>	C appears; cross-hairs disappear
21.	1 <r>	single bond between atoms 7 and 16
22. leave on atom 16	sp <r>	
23. position of atom 17	S <r>	S appears; cross-hairs disappear
24.	1 <r>	single bond between atoms 16 and 17

At this point the screen should look like Figure 7. To experiment with the other encoding directives, type D to exit the connection section and get the ENTER DIRECTIVES prompt again. Type 2 for the RING directive. The ring structures are automatically located and the ring atoms are numbered as in example 1.

Now suppose that you have made some errors in entering the structure. For example, suppose that atom 8 should be oxygen instead of carbon and that the bond between atoms 16 and 17 should be double instead of single. To correct these mistakes, type 4 for the ALTER directive. Next, you must enter another directive to indicate what kind of alternations are to be made. Since these are connections errors, type 1 for the connection directive. To change the type of atom 8 to oxygen:

1. Position the cross-hairs on atom 9
2. Type in the correct atom type (sp <r>)
3. Position the cross-hairs on atom 8
4. Type in the correct atom type (O <r>)
5. Cross-hairs will disappear; type 1 <r> to indicate single bond

No changes will appear on the screen yet, but they will be stored internally. Now to change bond type:

1. Position the cross-hairs on atom 16
2. Type in the correct atom type (sp <r>)
3. Position the cross-hairs on atom 17
4. Type in the correct atom type (S <r>)
5. Type in the correct bond type (2 <r>)

To check that the structure has been correctly altered and to exit the alter section, type @. The changes you have made should appear in the redrawn structure. For more information on making alternations, see Section 3.

To see how the stereochemistry section works, type 3 for the STEREO directive. The prompt

ASYMMETRIC ATOMS ?

appears. In this example, atom 9 is asymmetric. To indicate this to the program, position the cross-hairs on atom 9 and type A. The cross-hairs disappear and the prompt

ABOVE OR BELOW Z=0 ?

appears. Type BE for below. If there are four connections to the asymmetric atom you specify, one of the four connected atoms must be specified to be either above or below the Z=0 plane along with the asymmetric atom. Since there are four connections to atom 9, the prompt

FOURTH ATTACHMENT

and the cross-hairs appears. Position the cross-hairs on atom 10 and type sp <r> to indicate that this is the fourth attachment. There is no immediate visual feedback from the stereochemistry section, but the Z coordinates of the asymmetric atom and the fourth attachment have been changed internally. Type D to exit the stereochemistry section. Now you can type @ to have the molecular structure redrawn and view the changes. The structure should now look like Figure 8. For more information on the stereochemistry section of SFILES, see Section 3.

The structure has now been completed. To exit the ADD section, type 5 to enter the finish directive. Now type <r> to get the labelling prompts. In response to the message

5-FINISH

4-ALTER

3-STEREO

2-RING

1-CONNECTION
ENTER DIRECTIVE

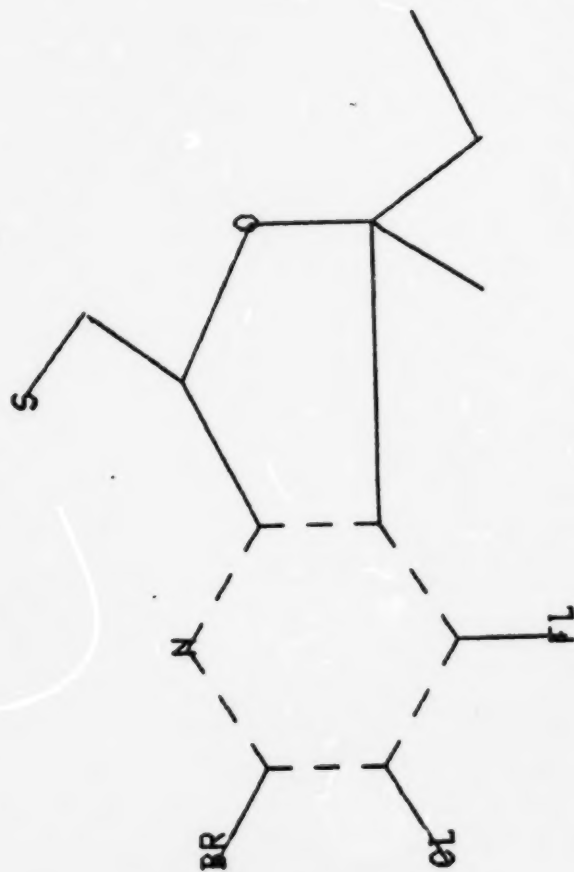


Figure 8. Final Structure Entered in Example 2.

INPUT ASSOC. INFORMATION

you can enter zero. This prompt will allow you to input up to five numbers and is normally intended for recording vapor pressure, etc. This data is not used in computations, it is strictly for labelling purposes.

This completes the examples on adding structures to the SFILES disk files. To delete a structure from the files that you no longer want, you must use the DELETE command. When you enter this command you will be asked for the library access number of the structure that you want to delete. Remember that to delete a structure of type "backbone" from the disk files you must have selected structure type backbone previously (i.e., either on entry to SFILES or by using the ALTER command). Of course, the same thing applies to the other structure types.

To avoid deleting the wrong structure by mistake, it is recommended that you use the REVIEW command to view the structure before deleting it in order to verify that it is really the one you want to wipe out.

3. GENERAL INFORMATION

3.1 Checking File Contents

There are three commands that provide information about file contents: CHECK, LIST, and LABEL. When you issue the CHECK command, SFILES outputs a report of the status of the structure, substructure, or backbone (depending on what mode you are in) file to the terminal in the format below:

```
NUMBER OF STRUCTURES  505
LARGEST LIB. ACC. NO.  877
1ST EMPTY POS IN MAIN FILE  7582
1ST EMPTY POS IN GEO FILE  2391
```

In this example, the structure file contains 505 structures and the largest library access number used is 877. There are 7582 records used in the topological information file, and 2391 records used in the geometrical information file.

```

TASK ? : LIST
NUMBER OF STRUCTURES 19
LARGEST LIB. ACC. NO. 19

LIST ALL LABEL ? : YES

ENTER LOWEST AND HIGHEST DAN TO BE CHECKED : 1 19

PRINT OUT LIST OF STRUCTURE : NO
1 BIPH BIPHENYL
2 BP02 PP'DICL BIPH
3 BP02 PENTA CL BIPH
4 T1
5 T2
6
7
8
9
10
11
12
13
14
15
16
17
18
19
NUMBER FOUND 19

TASK ? :

R001 BISPENOL A C15H16O2
R002 DBDBECH C8H12BR4
R003 IPBP C15H16
R004 DIPBP C18H22
R005 DCBTF C7H3CL2F3
R006 2-ETHYLHEXANOIC ACID
ETHER
TCDD
TUYT P-CHLOROPHENOL
EXAMPLE 1

```

Figure 9. Examples of LIST commands.

The CHECK command can be used to find out how much storage is used up by a specific structure by issuing the CHECK command immediately before and after entering the structure and noting the differences in the report numbers.

The capacity of each type of file is given below:

	<u>Structures</u>	<u>Substructures</u>	<u>Backbones</u>
max no. of structures	100	100	20
largest valid access no.	100	100	20
max no. of records in main file	14000	300	300
max no. of records in geo file	6000	100	150

The LIST and LABEL commands provide information about the file contents based on the labels that are assigned to structures when they are stored in the files.

When you issue the LIST command, the same information generated by the CHECK command is displayed as above. This is followed by the prompt

LIST ENTRIES WITH SUPPLIED LABEL ?

If you respond affirmatively, you are asked to specify the label and a range of library access numbers which will be searched for under the given label. The program then displays the number of structures that it found with that label. Sample output from the LIST command is illustrated in Figure 9.

When you issue the LABEL command, the prompt

LIST ALL LABELS ?

appears. If you respond affirmatively, you are asked to specify a range of library access numbers to be examined. Figure 10 represents sample output from the LABEL command:

TASK ? : LABEL

ENTER LOWEST AND HIGHEST DAN TO BE CHECKED : 1 19

LIST ON LP ? : NO

1	BIPH	1
2	BP02	2
3		9
4		1
10	R001	1
11	R002	1
12	R003	1
13	R004	1
14	R005	1
15	R006	1
18	TUYT	1

TASK ? :

Figure 10. Examples of LABEL command.

3.2 Adding and Changing Structures

The CHANGE command and the ADD command (when starting with a structure or backbone) are very similar. The only difference is that when you have completed the modifications of the structure under the CHANGE command, it is stored in the same location from which it was retrieved. Under the ADD command you must supply a new library access number where the structure will be stored.

The following information applies to structures drawn under either the CHANGE or ADD commands.

3.2.1 Alterations

To change any structural information that has been encoded, the ALTER directive must be executed. This is done by typing 4 in response to the ENTER DIRECTIVE prompt. You must then enter another directive number to indicate what kind of alterations are to be made. Connection type alterations such as deleting atoms and bonds or changing their types can be made by entering 1 for the connection directive. The cross-hairs appear on the screen. To change a bond or atom type, use the following procedure:

1. Position the cross-hairs on one of the atoms forming the bond.
2. Enter the correct atom type.
3. Position the cross-hairs on the other atom forming the bond.
4. Enter the correct atom type.
5. Enter the correct bond type.

To delete a bond, use the above procedure but enter a bond type of zero. Note that in the ALTER section, bonds can be deleted but not added. To add a bond, you must return to the main directive level and enter the connection directive.

To delete an atom from the structure, position the cross-hairs over the atom and type a slash (/).

The changes you make in the ALTER section are not immediately apparent, but the changes are stored internally. To view the modified structure, you must exit the ALTER section. This can be done in the following two ways:

1. Type D - this allows you to enter a new directive under the ALTER section. You can type 5 to return to the main directive level or you can type 2 or 3 to do alterations in the ring or stereochemical information.
2. Type @ - the structure is redrawn (including alterations) and you return to the main directive level.

Alterations in ring and stereochemical information are described in the Sections 3.2.3 and 3.2.4.

3.2.2 Entry of Charge Information

You have to enter charge information through the ALTER section of the drawing routines. The ALTER section is entered by typing 4 in response to the ENTER DIRECTIVE prompt. After entering the alter section, you must enter another directive to indicate the type of alteration to be made. Type 6 to enter charge information.

When the cross-hairs appear, position them on the atom that is to carry a charge. For negative charge, type a minus sign. For positive charge, type a plus sign. To remove a charge that was previously entered, type U. After the charge has been assigned, the ENTER DIRECTIVE prompt is reissued. To apply charge to another atom, enter 6 and repeat the process. To exit the ALTER section, follow the procedure described in Section 3.2.1.

3.2.3 Ring Information

Ring information for structures that contain completed rings are perceived by the program either when an explicit ring directive is given by the user or when the finish directive is issued. The ring section of the program finds the smallest set of basis rings which defines all of the ring systems in the structure.

After the ring section is executed, numbers appear next to each atom that is recognized as a ring member. This number represents the number of basis rings in which the atom is considered to be a member. However, on subsequent redrawing of the structure (e.g. under the REVIEW command), the numbers that appear next to each ring atom represent the size of each ring in which the atom is a member. Hence there will be more than one number

next to an atom if it is a member of more than one ring. - If the ring size is greater than ten, only the last digit will appear.

When entering a substructure in which you want to have ring atoms that are not contained in completed rings, you must take special action so that the program can recognize these atoms as ring members. After you have drawn the structure, type 2 in response to the ENTER DIRECTIVE prompt to go to the ring section. You must perform this step even though the ring section cannot locate rings that are not completed. Next, type 4 in response to the ENTER DIRECTIVE prompt to go to the ALTER section. In this section, you can add ring information explicitly as follows. Type 2 to go to the ring section under ALTER mode. The cross-hairs appear. To identify each ring atom (not already located by the ring section) as a ring atom:

1. Position the cross-hairs on the atom
2. Type R <r> to identify it as a ring atom
3. Enter the number of rings of which the atom is a member
4. Enter the sizes of the rings
5. The cross-hairs reappear; either position them on the next atom or type D to exit the ring section (you will still be under ALTER mode)

3.2.4 Specification of Stereochemistry

The double bond stereochemistry calculations of the program are completely automatic. After a structure has been encoded and the finish directive has been issued, the stereochemistry section of the program is executed. Visual feedback from this section occurs when a double bond is found to be either cis or trans. In this case, either a C or T appears in the center of the bond.

Information about asymmetric atoms must be entered explicitly by the user through the STEREO directive. This is done by typing 3 in response to the ENTER DIRECTIVE prompt. On entry to the stereo section, the cross-hairs appear along with the prompt

ASYMMETRIC ATOMS ?

Position the cross-hairs on an asymmetric atom and type A. The prompt

ABOVE OR BELOW Z=0 ?

appears. Type AB for above or BE for below. If there are less than four connections to the atom, the ASYMMETRIC ATOMS? prompt is reissued and you either designate the next asymmetric atom or type D to exit the stereo section. If there are four connections to the atom, one of the four connections must be specified to be either above or below Z=0 along with the asymmetric atom. In this case, the prompt

FOURTH ATTACHMENT ?

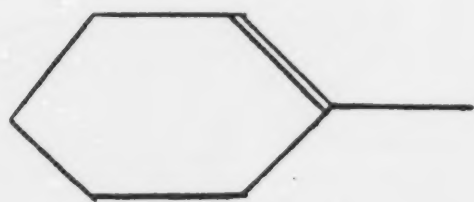
appears. To identify the fourth attachment, position the cross-hairs on the appropriate atom and type sp <r>. No changes in the structure are immediately visible, but the changes in the Z coordinates have been stored internally. When the structure is redrawn later, the changes will appear.

Alteration of stereochemical information is done under the ALTER section. After entering the alter directive, type 3 to enter the stereo section. To change a double bond, position the cross-hairs on the bond and type sp <r>. The specification is switched. To delete an asymmetric atom, position the cross-hairs on the atom and type A.

3.2.5 Drawing Structures Containing Double Bonds

When entering structures containing carbon-carbon double bonds, you must be careful to enter the surrounding atoms in the proper relationship to the atoms forming the double bond. The program perceives the geometry of the torsional angles about the double bond as they have been entered. Thus, if you enter an acyclic double bond in the cis configuration, it will always remain in this configuration during modelling. This also applies to the trans configuration. If you accidentally enter a double bond as trans within a ring, the modelling routine will never be able to generate a good model.

Also, remember that structures entered through SFILES must be represented in two-dimensional form. For example, a substituted cyclohexene should be entered as the drawing on the left in Figure 11, not as the drawing on the right:

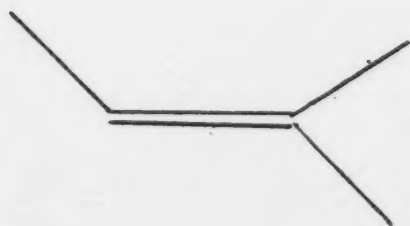


Correct

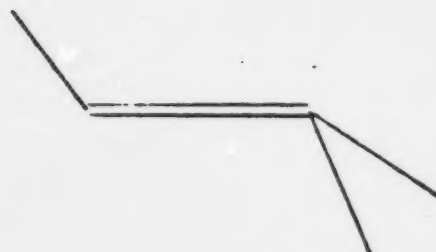


- Incorrect

and 2-methyl-2-butene should be entered as on the left, not the right:



Correct



Incorrect

Figure 11. Example of structure with double bonds.

3.3 Symbols for Atom/Bond Types

<u>Atom Type</u>	<u>Entry Symbol</u>
Carbon	sp (space)
Oxygen	O
Nitrogen	N
Sulfur	S
Flourine	F
Chlorine	C
Bromine	B
Iodine	I
Phosphorous	P
Silicon	T
Boron	U
Selenium	V
Arsenic	W
Unknown (substructures)	X
Unknown (structures)	Y
metal atom	M

<u>Bond Type</u>	<u>Entry Symbol</u>
single	1
double	2
triple	3
aromatic	4
de-localized	5
*none	0

* for deleting bonds in the ALTER section

3.4 Reasons for Bell Ringing

The bell serves as a signal for invalid user input during the program. The following list may give specific clues about an error, depending on what section of the program you are executing.

1. Reponse to the ENTER DIRECTIVE prompt
invalid directive number - must enter 1, 2, 3, 4, or 5

2. Connection section

- a) Unrecognizable element type
- b) Atom not located
- c) Invalid bond type

3. Stereochemistry section

- a) An A not entered to indicate an asymmetric atom
- b) Atom not found

4. Alter section

- a) Next directive number not entered
- b) Connection section - same as above
- c) Ring section
 - i. An "R" for ring atom not entered
 - ii. Atom not found
 - iii. Number of rings of which the atom is a member is greater than five
- d) Stereochemistry
 - i. Same as above
 - ii. Double bond not located
 - iii. Bond indicated is not double
 - iv. Atom not specified as asymmetric

4.0 SUMMARY OF SFILES COMMANDS

Prompt:

ENTER STRUCTURE TYPE

Possible responses:

- 1-STRUCTURE - The structure to be created or studied will be a complete chemical structure. A maximum of 1000 structures can be stored in the structure file.
- 2-SUBSTRUCTURE - The structure to be studied will be a chemical substructure. A maximum of 100 substructures can be saved in the substructure file.

3-BACKBONE - The structure to be studied will be a chemical backbone.
A maximum of 20 backbones can be saved in the backbone
file.

Prompt:

TASK?

Possible responses:

- ADD - The program enters the routine for adding a structure to the
structure files. Within the ADD routine there are several commands
available to the user.
- @ - If this command character is entered while drawing a structure will
cause the structure to be immediately redrawn.
- D - This command character indicates to the routine that the user is
finished drawing a structure and that he wishes to enter a new
directive.

Responses to the prompt:

ENTER DIRECTIVE

- 1 - This command indicates that the user wishes to connect atoms and add
new atoms in a structure. The entry symbols for atoms are:

Carbon	"space bar"
Oxygen	"O"
Nitrogen	"N"
Sulfur	"S"
Flourine	"F"
Chlorine	"C"
Bromine	"B"
Iodine	"I"
Phosphorous	"P"
Silicon	"T"
Boron	"U"
Selenium	"V"
Arsenic	"W"
Unknown ss	"X"
Unknown s	"Y"
Metal atom	"M"

The entry symbol for bonds are:

Single	"1"
Double	"2"
Triple	"3"
Aromatic	"4"

- 2 - This command indicates that the user wishes to have the ring routine identify the rings in the structure being studied. The ring routine is automatically called when the user enters the "5" command described below.
- 3 - This command is used when the user wishes to indicate the stereo-chemistry of a structure. The user is expected to indicate asymmetric carbon atoms by using the cross hairs and entering the letter "A". After each asymmetric atom is identified, the user must indicate whether the atom is above or below the plane by entering "AB" or "BE". For an asymmetric atom with four connections, one of the four connected atoms must also lie in the plane with the asymmetric atom. This atom is indicated by using the cross hairs and pressing the space bar followed by return. To exit the stereo directive use either the "@" or "D" commands described above. The stereo-chemistry of double bonds is automatically assigned by the routine after entering the "5" command described below.
- 4 - This command character indicates that the user wishes to alter the structure being entered. After entering this command the user must enter a "1" or "3" command described above. To delete an atom the entry symbol is "/". To delete a bond the entry symbol is "0".
- 5 - This is the finish command. After entering this command the routine will assign the ring information and the double bond stereo-chemistry. When the routine has finished making modifications to the structure the user should press the space bar and return keys. The user will then be prompted for a name, associated information, a four character label, and a library access (DAN) number.

Further responses to task:

- ALTER - This command indicates that the user wishes to change the type of structure being studied. The program will prompt the user with ENTER STRUCTURE TYPE as described above.
- CHANGE - This command allows the user to modify a structure or other information already saved in a structure file.

- CHECK - This command is used to list a summary of the contents of the structure files.
- DANM - This command indicates to the program that the user wishes to have each new DAN number automatically incremented.
- DELETE - This command is used to delete a structure from the structure files.
- DONE - This command is used to exit the SFILES program.
- HELP - This command prints a list of the available responses to the TASK prompt.
- INIT - This command initializes new structure files. This command should not be used since it will make all structures presently stored inaccessible to the SFILES and CLOGP programs.
- LABEL - This command allows the user to obtain a list of DAN numbers, labels, and names of compounds stored in the structure files.
- LIST - This command allows the user the DAN number, label, name, and associated information for a specific label or all labels of compounds stored in the structure files.
- REVIEW - This command allows the user to display any structure stored in the structure files.
- SMILE - This command activates structure entry via SMILES rotation. SMILES is the default structure entry method whenever a non Tektronix terminal is being used.
- NOSMILE - This command deactivates structure entry via SMILES rotation. Graphical entry via a Tektronix terminal is assumed.

APPENDIX A

This appendix describes the SMILES method of unambiguous molecular structure entry into the SFILES structure files. SMILES is a non-graphical, non-terminal specific structure entry method which may be accessed through SFILES in two ways. First, SMILES is automatically invoked in response to the ADD command of SFILES if the user has indicated that he is not using a Tektronix graphics terminal. Second, the SMILES mode can be activated by using the SMILES command in response to the SFILES Task prompt. Chemical structures entered via the SMILES method do not contain any graphical or stereochemical information and can only be reviewed as a structure connector table. Another limitation of structures entered via SMILES is that SMILES structures can not be added on to existing stored structures nor can SMILES structures be used as a starting point for a graphical structure.

In the following sections of this appendix the rules for SMILES notation will be described and examples of structure entry using SMILES will be given.

RULES FOR SFILES IMPLEMENTATION OF SMILES CHEMICAL NOTATION

The following rules allow the graph of a molecular structure to be unambiguously specified in SMILES notation.

1. Atoms are represented by their atomic symbols. The first letter of the atomic symbol must be uppercase, except that aromatic atoms may be specified by using lower case symbols. SMILES expects hydrogen-suppressed graphs and ignore or reject H. Valid atomic symbols are BR, C, CL, F, I, N, O, P, and S.
2. Bonds are represented by characters which separate atomic symbols; single bond is "-", double bond "=", triple bond "#", de-localized bond "=" and if needed the aromatic bond is ":" or ";" (see note below). The single bond character is optional; adjacent atoms are assumed to be singly bonded. If adjacent atoms are both aromatic (lower case), then an aromatic bond is assumed to join them.
3. A branch is indicated by enclosing it in parentheses (). Redundant parentheses are valid and optional. Nesting to 10 levels is allowed.

4. Ring closure is represented by the same single digit (in the range 1 to 9) immediately following each atomic symbol connected. Multiple ring closures to a single atom are valid.
5. The SMILES notation ends at the occurrence of the first space. Characters following the first space are considered text and may be used for a name or other description.
6. Formal charges are indicated by a signed integer enclosed in square brackets following the atomic symbol. It is not necessary to specify occurrence of unusual valences for hydrogen-suppressed graph generation. However, unusual valences may be indicated by an unsigned integer which shows the correct number of hydrogens attached (also within square brackets). If both formal charge and hydrogen count are specified, a comma separator is optional.

NOTE: Using ":" or ";" to define an aromatic bond is not exactly the same as using lower case letters for aromatic atoms. If aromatic atoms are defined by lower case letters all bonds between aromatic atoms are aromatic unless specifically indicated. The two methods of enter p-chlorobiphenyl are:

c1:c:c:c:c:c:l:c2:c:c:c(CL):c:c:2
and
clcccc1-c2ccc(CL)cc2

Examples using the GEMS, SFILES implementation of SMILES

6-hydroxy-1,4 hexadiene $C_6H_{10}O$
C=CCC=CCO

3-chloro-N-ethyl-butamide $C_6H_{12}NOCl$
CCC(CL)C(=O)NCC

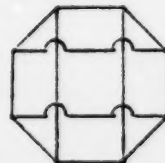
3-propyl-4-isopropyl-1-heptene $C_{13}H_{26}$
C=CC(CCC)C(C(C)C)CCC

2-methyl-4-chloro-2-cyclohexene $C_7H_{11}Cl$
C1C(C)=CC(CL)CC1

Cubane (Pentacyclo octane) C_8H_8
C1C2C3C4C2C1C4C31

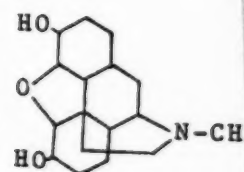
benzoic acid $C_7H_6O_2$

cl(C(=O)O)cccc1 or C1(C(=O)O);C;C;C;C;1



morphine $C_{17}H_{19}NO_3$

C1CC(N(C)CC3)C2C=CC(=O)C4C3C;1;C(=O);C(=O);C;C;1



SMILES INTERPRETER STANDARDS

SMILES interpreters have been implemented using FORTRAN, PASCAL, APL, and BASIC at various laboratories on a wide variety of machines. The following standards are offered to allow SMILES users a standard environment while allowing as much flexibility as possible for system development.

1. Node symbols must be combinations of one or more of the 52 upper and lower case English letters. Symbols and lengths are not specified in the general standard, but for chemical applications the use of atomic symbols is strongly recommended. The basic set is BR, C, CL, F, I, N, O, P, and S. Abridgements and extensions of this set must be documented. Recognition and warning of invalid symbol occurrence is required. Symbols representing multiple graph nodes are forbidden. The only interpreted use of letters will be for node symbols.
2. SMILES interpreters must recognize at least three edge types; single, double, and triple bonds are assigned the symbols -, =, and #, respectively. Absence of an edge descriptor will be interpreted as single bond, except that this requirement may be relaxed if edge descriptors are always required. (Although not recommended, this allows simple interpretation of all atomic symbols.)
3. The SMILES input string is scanned left to right and interpreted as a tree rooted at the first symbol. Branching is indicated by enclosing parentheses (). Nesting up to 10 deep must be interpreted correctly.
4. Ring closures are recognized by paired single digits following the node symbol at each end (and preceding the next bond descriptor). (For the GEMS, SFILES implementation if a : or ; is used as an aromatic bond descriptor it must occur before the ring closure digit.) Recognition of the numerals 1-9 is required for this purpose; (the numeral 0 may be used as well)*. Correct interpretation of nodes with multiple ring closures is required. Ring closures are assumed to be single or aromatic bonds; never double or triple (the GEMS, SFILES implementation requires the use of a ring closure bond descriptor).
5. Provision for direct input of aromaticity is required. This may be done by allowing the aromatic bond symbol ":", by recognizing lower case (first letter) node symbols as aromatic atoms, by both, or by an equivalent method.

6. Input to the SMILES interpreter is a character string (or equivalent byte array). The interpreter must accomodate a string of at least 80 characters. The input string may be passed to the interpreter or input within it. SMILES notation ends at the occurrence of the first blank.
7. Output must include: a) The number of nodes detected - zero for invalid or null input. b) A connection table or matrix. c) A bond table with numeric values, 1, 2, 3, 4, and 5 for single, double, triple, aromatic and de-localized words, respectively. d) *A node list with numeric

* Note allowed in GEMS, SFILES implementation

values corresponding to atomic number. Other data structures containing equivalent information are permissible replacements, for instance b, c, and d could be combined in an integer matrix. A node list of symbols, however, cannot replace the atomic number node list, but can of course be additionally provided.

8. *Interpretation of node modifiers is an optional SMILES extension. Node modifiers are placed in brackets [] following the node symbol and preceding the next edge symbol. When used, a signed digit indicates formal charge and an unsigned digit the hydrogen count. Commas are ignored within brackets. Letters are normally ignored within brackets; this may be relaxed to allow descriptions of various isomerisms to be introduced. Such nonstandard node extensions should be restricted to within the bracket node modifier.
9. *Fragment ends may be interpreted. If so, the asterisk represents required ex-substructure connections, and the question mark optional connections. Conventions for the specification of connection type are not standardized, although it is recommended that if needed, allowable bond types for each connection be interpreted inside unique enclosing symbols (such as <>) following the fragment end symbol.

* Not allowed in GEMS, SFILES implementation

NOTES ON SMILES RULES

The primary object of SMILES is to explicitly describe the hydrogen-suppressed graph of a chemical structure. Although information about hydrogen attachment is not directly specified, a simple convention allows the hydrogen-complete graph to be generated from a hydrogen-suppressed SMILES. This is to assume that hydrogen make up the remainder of an atom's lowest normal valence consistent with the explicit bond specification. Thus -C- is interpreted as -CH2- and -O as -OH. Using this convention, it is important to specify double bonds connecting both oxygens in fragments such as nitro- and sulfonyl- groups. An additional benefit resulting from this convention is that the symmetry of such oxygens is correctly represented. "Lowest normal valence" is usually taken to mean 3 or 5 for phosphorous and 2, 4, or 6 for sulfur. This insures, for example, that O = P(O)(O)O is interpreted as H3PO4 (orthophosphoric acid) while P alone represents phosphine (PH3). Note that Rule 7 permits specification of structures containing atoms of unusual valences without loss of generality.

Definition of an aromatic bond type is required for accurate specification of molecular structure.

It is recognized that various chemical information uses require differing definitions of aromaticity. NMR spectroscopy, synthetic chemistry, and chemical nomenclature each use the idea of aromaticity in a different context. By allowing the user to specify aromatic atoms and bonds, SMILES should be capable of serving most uses. Indeed, many purposes do not require accurate specification of aromaticity: it is perfectly valid in these cases to specify any kekule' structure, using single and double bonds. The resultant SMILES interpretation will be correct in every respect except for a loss of some symmetry information.